論文

ファジィ近接関係を用いた
動詞シソーラスの構築[†]藤本 太郎^{*1} 菅野 道夫^{*2}

自然言語処理を行う上でシソーラスは不可欠であるが、その構築には多くの労力が必要であった。そこで、相互情報量に基づく関連度を用いたシソーラス構築の自動化手法などが提案されてきた。しかし、相互情報量を用いた関連度の計算は複雑であり、多数のシソーラスを構築する必要がある場合などには高い計算コストが問題になっていた。

そこで、本研究では低い計算コストで相互情報量に基づく手法より高精度なシソーラス構築システムを構築するために、動詞間のファジィ近接関係を用いた関連度を用いたシソーラス構築システムを提案する。コーパスの出現頻度情報を反映させるために低頻度動詞の影響を排除するファジィフィルタと、高頻度名詞の影響を排除するファジィフィルタを提案し、ファジィ近接関係と合わせて4通りの関連度を定義した。

対照実験として相互情報量を用いた関連度と本手法を比較する。また、ファジィフィルタの効果を評価するために相互情報量を用いた関連度にファジィフィルタを適用した関連度も定義し、さらに4種類の関連度を定義した。EDR 共起辞書から名詞と動詞の関係を共起に関わる助詞「を、に、が、は、で」毎に5種類の分類実験に用いる共起データを作成した。5種類のデータと8種類の関連度の組み合わせで40通りの関連度を定義し、分類語彙表に登録されている動詞の再分類を行うことで評価を行った。さらに、本手法を用いて分類語彙表に未登録の動詞を分類し、人間が評価を行った。

両方の実験の結果、助詞「で」によって共起するデータをファジィ近接関係に低頻度動詞に関するファジィフィルタを適用した関連度が最も高い評価を得た。また、データ量を変化させて計算時間を計測し、データ量が多いほど本手法による計算量低減効果が高くなることを示した。以上より、従来手法に比べて低コストで高精度な動詞シソーラス構築が可能であることを示した。

1. はじめに

自然言語処理を行う上ではシソーラスは不可欠であり、そのシソーラスがシステムの性能を左右する。従来、シソーラス構築には多くの人手を要していたが、近年の電子化辞書[11]などの出現により計算機上でデータを扱うことが可能になり、シソーラス構築の自動化の研究が行われてきている。

シソーラスは単語の意味的な関係を表現したデータベースであり、分野毎にシソーラスを持つことによって単語の矛盾・多義の問題を緩和するなどが求められている[16]。

従来は相互情報量を用いたシソーラスの構築手法[1,2]などが提案されていた。しかし、計算が複雑であるため、多数のシソーラスを構築する場面には適さず、低い計算コストの手法が望まれる。

本研究では多様な分野毎のシソーラスの構築を容易

にするために、動詞シソーラスを従来と同等以上の性能を持ちながら低い計算コストで構築する手法としてファジィ近接関係を用いた動詞の分類手法を提案する。本研究では、名詞と5種類の格助詞を観点[9]として、「名詞-助詞-動詞」の組数をEDR 共起辞書[11]から求め、その頻度に基づいて動詞の関連度を定義する。本研究ではファジィ近接関係[6,3]を満たす関連度[17]に、低頻度動詞の影響を減ずるファジィフィルタ、高頻度名詞の影響を減ずるファジィフィルタを組み合わせ合計4種類の関連度を定義し、比較を行う。また、対照実験として従来手法である相互情報量[1,2]による関連度と比較を行う。ファジィフィルタの効果を示すために、相互情報量に対してもファジィフィルタを適用した関連度を定義した。

本研究で定義した関連度の評価として分類語彙表[12]に含まれる動詞を分類し、元の分類に含まれるかどうかを再現率、適合率で評価する。また、データ数を変化させて計算時間の比較を行う。さらに定義した関連度を用いて分類語彙表に未登録の動詞の分類を行い、アンケートによる評価を行う。

提案するファジィを用いた手法を用いることで、増加する語彙環境や多様な分野に対応する従来手法と同等以上の精度で低コストなシソーラス構築が可能になる。

本論文では2章において8種類の関連度を定義し、3章

[†] A Study on the Construction of Verb Thesauri using the Fuzzy Proximity Relation
Taro FUJIMOTO and Michio SUGENO

^{*1} 東京工業大学 大学院総合理工学研究科 システム科学専攻
(現 横浜国立大学 知能システム研究部)
Dept. of Systems Science, Tokyo Institute of Technology
(Intelligent Systems Laboratory, Fujitsu Laboratories Ltd.)

^{*2} 東京工業大学 大学院総合理工学研究科 システム科学専攻
Dept. of Systems Science, Tokyo Institute of Technology

で評価を行う。4章で本手法を用いて分類語彙表の拡張に適用した例を示し、5章において結論を述べる。

2. 関連度

本研究では、ファジィソーラス構築の際に用いる関連度[17]を用いた動詞間の関連度を提案し、従来からソーラス構築などに用いられる相互情報量に基づく関連度[14]と比較を行う。また、これらの関連度に対して低頻度動詞の影響を減ずるファジィフィルタと高頻度動詞の影響を減ずるファジィフィルタを提案する。2種類のファジィフィルタの組み合わせで4種類のフィルタが構成でき、2種類の関連度との組み合わせで合計8種類の関連度を定義する。

2.1 ファジィ近接関係を用いた関連度(pr)

本研究では要素間の曖昧さを含む関係を量的に記述するための手法であるファジィ関係を関連度としたソーラスの構築手法を提案する。

式1にファジィソーラス構築の際に用いる近接関係[17]を簡略化した関連度 pr を示す。名詞 $Y = y_1, y_2, \dots, y_n$ と動詞 $X = x_1, x_2, \dots, x_m$ の関係において、 x_i と共起する名詞の集合を R_{x_i} とする。この時、 x_i と x_j の間の関係の強さ pr を次式で定義する。

$$pr(x_i, x_j) = \frac{|R_{x_i} \cap R_{x_j}|}{\min(|R_{x_i}|, |R_{x_j}|)} \quad (1)$$

これは反射的、対称的であるため、ファジィ近接関係(fuzzy proximity relation)[6,3,7]を満足するが、ファジィ推移律を含む推移律を満たさない。また、この定義では推移律が成り立たないために、部分的に矛盾を含んでいるために不確実性が生じる[5]。そのためこの定義によって導かれる関係の強さはファジネスと考えられる。

例を挙げると、動詞 x_i と x_j において、 x_i は名詞 $\{y_1, y_2, y_3\}$ と共起し、 x_j は名詞 $\{y_2, y_3, y_4, y_5\}$ と共起していたと仮定すると、 $R_{x_i} = \{y_1, y_2, y_3\}$ 、 $R_{x_j} = \{y_2, y_3, y_4, y_5\}$ となる。よって x_i と x_j の関連度 $pr(x_i, x_j)$ は

$$\begin{aligned} pr(x_i, x_j) &= \frac{|R_{x_i} \cap R_{x_j}|}{\min(|R_{x_i}|, |R_{x_j}|)} \\ &= \frac{|\{y_2, y_3\}|}{|R_{x_i}|} = \frac{2}{3} \end{aligned}$$

となる。

2.2 相互情報量を用いた関連度(mi)

対照実験として、相互情報量を用いた名詞の分類の手法を用いて動詞の分類を行なう。ここでは新納[14]の定義を用いて、相互情報量を用いた関連度の求め方を示す。まず、動詞 x_i と名詞 y の相互情報量 $I(x_i, y)$

は次式で表される。

$$I(x_i, y) = \log_2 \frac{p(x_i, y)}{p(x_i)p(y)} \quad (2)$$

ここで $p(x_i)$ は x_i の出現確率、 $p(y)$ は y の出現確率、 $p(x_i, y)$ は x_i と y の同時出現確率を表している。このとき名詞 y_k から見た動詞 x_i と動詞 x_j の関連度 $sim(y_k, x_i, x_j)$ は次式で定義される。

$$\begin{aligned} sim(y_k, x_i, x_j) &= \begin{cases} \min(|I(y_k, x_i)|, |I(y_k, x_j)|), \\ \text{if } |I(y_k, x_i) \cdot I(y_k, x_j)| > 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

最終的に相互情報量を用いた動詞 x_i と動詞 x_j の関連度 $mi(x_i, x_j)$ は以下のように定義される。

$$mi(x_i, x_j) = \sum_k sim(y_k, x_i, x_j) \quad (4)$$

2.3 ファジィフィルタ

2.1節のファジィ関連度にコーパスの大域的な情報を反映させるために、低頻度動詞の影響を減少させるフィルタ及び高頻度名詞の影響を減少させるフィルタの2種類のファジィフィルタを定義する。

出現頻度はコーパスからの知識抽出の際に得られる重要な情報である。しかし、式1で定義したファジィ近接関係を用いた関連度にはコーパス全体における出現頻度が反映されていないため、本研究では大域的な出現情報を反映する手法としてファジィフィルタを定義する。

従来、出現頻度を関連度に反映する手法として

- ・出現頻度を定数に乗算する方法[13]、
- ・仕事量を基準にする方法[18]、
- ・相対的な出現頻度を用いる方法[15]、
- ・閾値以上のものを用いる方法[4]、

などがある。提案するファジィフィルタは、出現頻度を乗算する手法と閾値を用いて切り捨てる方法の折衷である。

一般に出現頻度とその出現頻度を持つ単語の関係は、出現頻度が3以下の単語がほとんどで、頻度が高くなるにつれて単語が急激に減少する。そのため、出現頻度を線形に重み付けすると、頻度が極端に高い単語があった場合に、その他の単語の重みが過度に小さくなる。他方、閾値で切り捨てる低頻度の単語の情報が失われてしまう。よって、本研究では両者の折衷であるファジィフィルタを定義する。このフィルタは「関連度計算に用いるデータの集合」を仮定した場合、あるデータがこの集合に含まれる度合い(グレード)を示していると考えられる。

二つの単語が出現する文脈が等しければ、それらの単語の意味が近いと考えられる[8]。そこで、動詞の出現頻度が低いという事は、等しい環境の文脈も少ないという事になり、その関連度の信頼性が低いということになる。他方、頻度が高い名詞は複数の意味に用いられることもあり得るため、同じ名詞で共起していたとしても異なる文脈である可能性があり、関連度の信頼性が低くなると考えられる。そこで、低頻度動詞の影響を減少させるフィルタと高頻度名詞の影響を減少させるフィルタを以下に定義する。

低頻度動詞の影響を減少させるフィルタ(lc) 動詞に関して出現頻度が平均以上ならば重みを1とし、平均未満ならば出現頻度が少なくなるにつれて重みを下げる。動詞の重み lc を以下のように定義する。

$$lc(x) = \begin{cases} 1 & (\text{If } x \geq \bar{x}) \\ x/\bar{x} & (\text{If } x < \bar{x}) \end{cases} \quad (5)$$

ただし、 x は動詞の出現数、 $\bar{x}$ は動詞の平均出現数である。

高頻度名詞の影響を減少させるフィルタ(hc) 名詞に関して出現頻度が平均以下ならば共有する1個の名詞の数の重みを1とし、出現頻度が平均以上ならば出現頻度が多くなるにつれて重みを下げる。名詞の重み hc を以下のように定義する。

$$hc(y) = \begin{cases} \bar{y}/y & (\text{If } y \geq \bar{y}) \\ 1 & (\text{If } y < \bar{y}) \end{cases} \quad (6)$$

ただし、 y は名詞の出現数、 $\bar{y}$ は名詞の平均出現数である。

2.4 8種類の関連度

2.1節のファジィ関連度及び2.2の相互情報量を用いた関連度に、2種類のファジィフィルタを組み合わせる以下に挙げる合計8種類の関連度を定義する。

- ・ファジィ関連度(pr) (式1で定義済み)
- ・ファジィ関連度に低頻度動詞に関するフィルタをかけたもの(pr_lc)

$$pr\_lc(x_i, x_j) = \frac{|R_{x_i} \cap R_{x_j}|}{\min\{|R_{x_i}|, |R_{x_j}|\}} \cdot lc(x_i) \cdot lc(x_j) \quad (7)$$

- ・ファジィ関連度に高頻度名詞に関するフィルタをかけたもの(pr_hc)

$$pr\_hc(x_i, x_j) = \frac{\sum_{y \in R_{x_i} \cap R_{x_j}} hc(y)}{\min\{|R_{x_i}|, |R_{x_j}|\}} \quad (8)$$

- ・ファジィ関連度にフィルタを両方かけたもの(pr_lchc)

$$pr\_lchc(x_i, x_j) = \frac{\sum_{y \in R_{x_i} \cap R_{x_j}} hc(y)}{\min\{|R_{x_i}|, |R_{x_j}|\}} \cdot lc(x_i) \cdot lc(x_j) \quad (9)$$

- ・相互情報量を用いた関連度(mi) (式4で定義済み)
- ・相互情報量を用いた関連度に低頻度動詞に関するフィルタをかけたもの(mi_lc)

$$mi\_lc(x_i, x_j) = mi(x_i, x_j) \cdot lc(x_i) \cdot lc(x_j) \quad (10)$$

- ・相互情報量を用いた関連度に高頻度名詞に関するフィルタをかけたもの(mi_hc)

$$mi\_hc(x_i, x_j) = mi(x_i, x_j) \cdot \frac{\sum_{y \in R_{x_i} \cap R_{x_j}} hc(y)}{|R_{x_i} \cap R_{x_j}|} \quad (11)$$

- ・相互情報量を用いた関連度にフィルタを両方かけたもの(mi_lchc)

$$mi\_lchc(x_i, x_j) = mi(x_i, x_j) \cdot \frac{\sum_{y \in R_{x_i} \cap R_{x_j}} hc(y)}{|R_{x_i} \cap R_{x_j}|} \cdot lc(x_i) \cdot lc(x_j) \quad (12)$$

3. 評価実験

2節で定義した8種類の関連度を用いて、EDR コーパス[11]と分類語彙表[12]からシソーラスを構築し、評価を行う。図1に示すように、コーパスから5種類の格助詞毎にデータを抽出し、それぞれのデータに対して8種類の関連度を用いて関係の強さを計算する。すなわち2つの動詞間に40通りの関連度を求めることになる。この関連度を用いて分類語彙表に登録されている動詞

図1 評価実験の流れ

が元の分類に分類されるかどうかを再現率・適合率によって評価する。

3.1 コーパスからのデータ抽出

本研究ではシソーラス構築のための頻度情報をEDR 共起辞書[11]から得る。頻度の高い5種類の格助詞「を、に、が、は、で」によって共起する「名詞+助詞+動詞」形式のデータ(表1参照)を抽出する。これら5つの助詞で全データの81.9%を占める。出現数が少ない格助詞は関連度を計算するために十分にデータが抽出できないと考え、5種類の助詞のみで評価を行う。この形式のデータを用いる理由は、助詞と動詞を一つの観点[9]としている手法に倣って、名詞と助詞を観点として動詞を分類するためである。抽出されたデータ数は表2に示す通りである。

3.2 分類

3.1節で抽出された5種類のデータ形式に関して、5節で定義された8種類の関連度を用いて動詞間の関連度を計算する。

この計算結果を用いて分類語彙表[12]に含まれる動詞を分類する。分類語彙表は表3に示すように、分類番号、分類見出し、同義語組で表される日本語シソーラスである。

分類の手順は以下に示す通りである。

1. あらかじめ分類語彙表の分類番号の深さをそろえるために、上位3桁の分類番号で揃える。表3の例では分類番号2.135のものを2.13に、2.1500のものを2.15に切り捨てを行う。
2. 動詞組のうち、両方の動詞が分類語彙表に登録されているものを抽出する。
3. ある動詞に関して、各分類に含まれる動詞の中で関連度が0より大きいものの平均が最も大きい分類がその動詞の分類とする。ここで平均を用いた理由は動詞のある分類に対する帰属度を正規化するためである。

4. 動詞が元の分類に分類された場合を正解、そうでない場合を不正解として各データと関連度毎に再現率、適合率を求める。

表1 EDR 共起辞書より抜きだした名詞+助詞+動詞データの例

名詞	助詞	動詞
共振回路	は	実現
共振回路	を	接続
共振回路	を	作
強心剤	を	投与
共振周波数	が	決ま
共振周波数	が	変化

表2 出現数上位5助詞とその助詞に関する名詞-動詞組の数

助詞	組数	助詞	組数
を	114425	に	83346
が	71998	は	70534
で	33366		

図2 各助詞、関連度毎に得られた分類の再現率

表3 分類語彙表の一部

分類番号	分類見出し	段落番号	同義語組
2.135	障り	1	さわる さしつかえる さしさわる
		2	さしくる くりあわせる
		3	あく
2.14	力	1	力む 激する 勢い込む
		2	あおる 掻き立てる 盛り上がる 盛り上げる 盛り返す
		3	ひしめく はりあう
2.1500	変化	1	変える 変わる 変ずる 移り変わる 変わり果てる 移ろう
		2	化する ばける なり果てる なりゆく 帰する
		3	入れ違ふ 入れ違える

3.3 評価

本節では、5種類の助詞及び8種類の関連度間の適合率・再現率、計算量の3点から評価を行う。ここで用いる再現率、適合率は式13、式14のように定義される。

$$\text{再現率} = \frac{|A \cap B|}{|B|}$$

$$\text{適合率} = \frac{|A \cap B|}{|A|}$$

ただし、 A =システムの解答の集合、 B =テストデータの解答の集合とする。3.2で得られた8つの関連度及び5種類のデータに関する適合率と再現率を付録Aの表6、表7に示す。これをグラフ化したものが図2、図3である。図4にデータ量と計算時間及び速度向上率を示す。

5種類の助詞の比較 表6、表7及び図2、図3に示すように、5種類の助詞を比較すると、再現率・適合率とも「は」が最も低く、「で」が最も高い値である。これは「は」が提題助詞であり、他の格助詞が固定された用法でしか用いられないのに対し、主題を提示する役割を持つため[10]、他の格助詞と比較して分類精度が低くなったと考えられる。「で」は出現数も少なく、格助詞としても限定された用法を持つため、分類精度が高いと考えられる。

8種類の関連度 表6、表7及び図2、図3に示すように、関連度 pr と mi を比較すると、すべての助詞に対して再現率・適合率とも mi の方が高い。これはコーパス全体の情報を用いている mi は、2動詞とその共起する名詞情報のみしか用いない pr より正確にシソーラス構築を行うことができることを示している。

pr に対して lc 、 hc のいずれのファジィフィルタを加えても再現率・適合率は上昇するが、 lc の方がより効果的である。両方を加えた場合は lc を加えた場合よりも再現率・適合率とも低い値をとる。このことは、頻度が低い動詞の数より頻度が高い名詞の数が多いため、効果に差が出たと考えられる。

mi に対しては片方のフィルタでは再現率・適合率とも悪化している。式2の相互情報量にファジィフィルタ lc を適用すると、低頻度の動詞に関して逆の動作をして効果を打ち消しあって関連度性能が低下していると考えられる。 hc を適用した場合は高頻度名詞の影響を弱め合っていると考えられる。

以上の議論から、最も再現率、適合率が高い値を示すのは助詞「で」によって共起されたデータを pr_lc を関連度として用いて分類を行った場合である。

計算量 表4は助詞「で」を含む共起データのデータ数

図3 各助詞、関連度毎に得られた分類の適合率

図4 データ数と計算時間

表4 関連度と計算時間の関係(単位: 秒)

データ数	mi	pr_lc	速度向上率
500	6.53	6.31	3.4%
1000	26.25	25.36	3.4%
5000	348.27	334.55	3.9%
10000	1273.40	1194.44	6.2%
29147	6082.02	5617.45	7.6%

を変化させて計算時間を測定した5回の平均をとったものと mi を基準として pr_lc がどれだけ高速化を実現したかを速度向上率として示したものである。これをグラフ化したものが図4である。データ量の最大値29147は助詞「で」を含むデータの最大量である。すべてのデータ数で3.4%から7.6%の速度向上が見られ、データ数が増えるほど速度向上率が上昇している。

データ数が増加することで速度が向上する理由を、 mi および pr_lc を計算する際に、データ数に比例して計算時間が増える部分と単語の組合わせに比例して

計算時間が増える部分の計算量に分けて考察する。

まず、データ量に比例して計算量が増える部分は *mi* の場合は式2に示す相互情報量、*pr_lc* の場合は式5に示す *lc* である。助詞「で」の29147データで5回計算を行って計算時間の平均をとると相互情報量は23.9秒、*lc* は4.0秒であった。

次に単語の組合わせに比例して計算量が増える部分は、*mi* の場合は式3、式4、*pr_lc* の場合は式1である。*mi* は446673関係を6061.63秒で計算し、1関係当り0.0136秒かかる。他方 *pr_lc* は546522関係を5613.45秒で計算し、1関係当り0.0103秒である。

単語の組合わせを計算する時間は等しいために、データ量が増加するにしたがって *mi* と *pr_lc* の計算時間の差が開いていくと結論づけられる。

4. 分類語彙表拡張への応用

3.2節で動詞間の関連度を計算して分類語彙表[12]に登録されている動詞の分類を行った方法を用いて、分類語彙表に登録されていない動詞についても分類を試みる。この分類に対し、27人から得たアンケートで評価を行う。

4.1 分類、評価手順

以下の手順で分類、評価を行う。

1. 節で用いた動詞間の関連度データのうち、片方だけの動詞が分類語彙表に含まれているものを抽出する。
2. 助詞「で」を持つデータから作成された関連度データから分類語彙表に分類されていない1067動詞を抽出する。
3. 各動詞と分類の関連度を計算する。関連度は分類中に含まれる全動詞と分類される動詞の関連度の平均を用いる。
4. 抽出された動詞の分類候補について、27人に対し、助詞「で」によって共起する名詞--動詞関係によって得られた分類候補の動詞からランダムに選んだ19単語についてアンケートを行なう。アンケートは各項目に対してその分類に含まれるか含まれないかを○×の2択で行なう。
5. その分類が正しいと見做した人が過半数の分類についてその単語はその分類に含まれると判断する。アンケートによって得た正答の例を表5に示す。人が行なった分類を正答として、適合率・再現率を求める。%ただし、アンケートにおいて複数の正答が得られる場合もある。

4.2 評価

4.1節で行った分類によって得られた動詞分類の一例を付録Cに示す。表5にも示すように、新たに分類された動詞には動詞化した名詞(サ変動詞)や複合動詞が多く含まれている。

また、付録Bの表8、表9に各データ、各関連度を用いて分類を行った評価を示す。ここで用いる再現率・適合率は式15、式16のように定義される。

$$\text{再現率} = \frac{|A \cap B|}{|B|} \quad (15)$$

$$\text{適合率} = \frac{|A \cap B|}{|A|} \quad (16)$$

ただし、*A*=システムの解答の集合、*B*=アンケートで得た回答の集合 とする。これをグラフ化したものが図5、図6である。

図5 各助詞、関連度毎に得られた分類の再現率

図6 各助詞、関連度毎に得られた分類の適合率

グラフ2とグラフ5、グラフ3とグラフ6をそれぞれ比較するとグラフの形式は異なっているが、分類語彙表に登録されている動詞の再分類と同様に、助詞「で」を含むデータを関連度 pr_lc で計算して分類を行ったデータの評価が最も高く、アンケートの結果を正解として評価した場合の再現率は0.349、適合率は0.033で最も高い評価であった。

値に関しては分類語彙表に登録されていた動詞を再分類した評価と比較して約1/2の値である。これは分類語彙表の分類と人間の分類が必ずしも一致しないことや、コーパスを学習に用い、データと評価に用いたデータが異なること、アンケートを行った動詞が少なく、偏りがあったなどの理由が考えられる。

本システムだけで自動シソーラス構築システムを構築することは精度が低いために困難であるが、システムが上位いくつかの分類候補とその関連度を提示しその中から選んでもらうなどの人間とのインタラクションを含んだ半自動構築システムを構成することが可能と考えられる。

5. おわりに

本研究では、ファジィ関連度を定義し、名詞と動詞の共起データから動詞間の関連度を計算し、分類を行った。

まず、関連度としてファジィ近接関係[17]を用いた関連度(pr)に、低頻度動詞の影響を減少させるファジィフィルタ(lc)と、高頻度名詞の影響を減少させるフィルタ(hc)の2種類のファジィフィルタを組み合わせ

表5 アンケートで得た答えの例

動詞	アンケートで得た答
理解する	2.30
表現する	2.31, 2.30
導入する	2.15
提供する	2.15
説明する	2.38, 2.30
実現する	2.15, 2.30
指摘する	2.30
発表する	2.30, 2.37, 2.31
入れる	2.15
回す	2.30

でかけることで4種類の関連度($pr, pr\_lc, pr\_hc, pr\_lchc$)を定義した。

対照実験として従来手法の相互情報量に基づく関連度(mi) [1,2]を用い、ファジィフィルタの特性を調べるために mi に適用したもの(mi_lc, mi_hc, mi_lchc)を定義した。EDR 共起辞書[11]から出現頻度が高い「を、に、が、は、で」の5種類の助詞によって共起する名詞と動詞の組を用いて分類実験を行った。

まず、8種類の関連度と、5種類のデータから40種類の動詞間の関連度を求め、分類語彙表[12]を元に動詞の分類を行った。助詞に関しては「で」が最も高い評価を得た。また、関連度に関してはファジィ近接関係を用いた関連度に低頻度動詞に関するファジィフィルタをかけたもの(pr_lc)が最も高い評価を得た。また、データ数を変化させて計算時間を比較したところ、 mi を基準として pr_lc は3.4%から7.6%高速であり、デ

A 分類語彙表に登録されている動詞を再分類した際の評価

表6 各助詞、関連度毎に得られた分類の再現率

	pr	pr_lc	pr_hc	pr_lchc	mi	mi_lc	mi_hc	mi_lchc	平均
を	0.481	0.539	0.490	0.505	0.444	0.215	0.216	0.505	0.424
に	0.344	0.530	0.524	0.489	0.472	0.134	0.134	0.489	0.390
が	0.498	0.579	0.586	0.560	0.546	0.158	0.160	0.560	0.456
は	0.369	0.522	0.470	0.496	0.441	0.154	0.162	0.496	0.389
で	0.541	0.593	0.571	0.565	0.513	0.220	0.236	0.565	0.476
平均	0.447	0.553	0.528	0.523	0.483	0.176	0.182	0.523	0.427

表7 各助詞、関連度毎に得られた分類の適合率

	pr	pr_lc	pr_hc	pr_lchc	mi	mi_lc	mi_hc	mi_lchc	平均
を	0.042	0.053	0.046	0.049	0.041	0.025	0.025	0.049	0.041
に	0.027	0.046	0.046	0.046	0.038	0.014	0.013	0.046	0.035
が	0.041	0.055	0.053	0.054	0.049	0.017	0.017	0.054	0.043
は	0.030	0.049	0.042	0.046	0.036	0.015	0.015	0.046	0.035
で	0.047	0.055	0.051	0.054	0.048	0.024	0.026	0.054	0.045
平均	0.037	0.052	0.048	0.050	0.042	0.019	0.019	0.050	0.040

B 分類語彙表に登録されていない動詞を分類した際の評価

表8 未登録の動詞を分類しアンケートで再現率を評価した結果

	pr	pr_lc	pr_hc	pr_lchc	mi	mi_lc	mi_hc	mi_lchc	平均
を	0.209	0.209	0.174	0.174	0.200	0.200	0.174	0.174	0.189
に	0.277	0.277	0.245	0.223	0.277	0.277	0.245	0.223	0.255
が	0.289	0.289	0.145	0.105	0.289	0.289	0.145	0.105	0.207
は	0.305	0.305	0.293	0.293	0.305	0.305	0.293	0.293	0.299
で	0.349	0.349	0.270	0.238	0.349	0.349	0.270	0.238	0.302
平均	0.286	0.286	0.225	0.207	0.284	0.284	0.225	0.207	0.250

表9 未登録の動詞を分類しアンケートで適合率を評価した結果

	pr	pr_lc	pr_hc	pr_lchc	mi	mi_lc	mi_hc	mi_lchc	平均
を	0.021	0.021	0.019	0.020	0.020	0.020	0.019	0.020	0.020
に	0.026	0.026	0.026	0.025	0.025	0.025	0.026	0.025	0.025
が	0.026	0.026	0.016	0.013	0.026	0.027	0.016	0.013	0.020
は	0.025	0.025	0.027	0.027	0.025	0.025	0.027	0.027	0.026
で	0.033	0.033	0.030	0.027	0.033	0.033	0.030	0.027	0.031
平均	0.026	0.026	0.024	0.022	0.026	0.026	0.024	0.022	0.025

C 分類語彙表に追加される動詞の候補

分類語彙表に未登録の動詞を分類し、得られた分類先の候補を示す。左列が動詞、中列が分類先の候補、右列が分類先の候補に含まれる動詞である。

表10 分類語彙表に追加される動詞の候補の例

動詞の候補	分類	動詞の例
ついて行ける	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
オープンする	2.11	対する 似る 着ける
	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
	2.37	持つ 支払う 儲ける 商う
	2.38	たがやす かもす 漕ぐ 仕立てる 設ける 作る
	2.58	生まれる 死ぬ こなれる 病む
安定する	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
	2.38	たがやす かもす 漕ぐ 仕立てる 設ける 作る
	2.58	生まれる 死ぬ こなれる 病む
位置づける	2.11	対する 似る 着ける
	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
	2.38	たがやす かもす 漕ぐ 仕立てる 設ける 作る
	2.58	生まれる 死ぬ こなれる 病む
位置合せする	2.12	ある 現われる 出来る 仕上がる 残る 保つ
	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
飲酒運転する	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
	2.38	たがやす かもす 漕ぐ 仕立てる 設ける 作る
運び込む	2.12	ある 現われる 出来る 仕上がる 残る 保つ
	2.15	変える 動く 移る 出る 上がる 合う 触れる 折る 殖える
	2.30	感ずる 呼ぶ 著わす はやる する 交わる 治まる 持つ たがやす
	2.38	たがやす かもす 漕ぐ 仕立てる 設ける 作る

ータ数が増加するほど高速化が実現できることを示した。

次に、分類語彙表に未登録の1067動詞を分類を試みた。アンケートによる評価では、動詞の再分類の結果と同じく、助詞「で」によって共起する名詞と動詞の組と *pr_lc* を用いた関連度が最も高い評価を得たが、評価値が十分に高くないために完全な自動化は困難である。しかし、シソーラスに追加される動詞の候補を提示し、人間が選択することによってシソーラスを拡張するインタラクティブなシステムに組込むことでシソーラス半自動構築に応用可能と考えられる。

謝辞

本研究を進める上で有用なコメントを頂いた東京工業大学知能システム科学専攻の内海彰講師及び富士通研究所ネットメディア研究センターの岡本青史氏に感謝します。

参考文献

- [1] R. Basili, M. T. Pazienza, and P. Velardi. A Context Driven Conceptual Clustering Method for Verb Classification, chapter 7, pages 117 {142. MIT Press, 1996.
- [2] R. Basili, P. Velardi, and M. T. Pazienza. Integrating general-purpose and corpus-based verb classification. Computational Linguistics, 22(4) : 559 {568, 1996.
- [3] D. Dubois and H. Prade. FUZZY SETS AND SYSTEMS. ACADEMIC PRESS, New York, 1980.
- [4] F. Fukumoto and J. Tsujii. Automatic recognition of verbal polysemy. Proc. 15th COLING, 1994.
- [5] J. G. Klir and T. A. Folger. 本多 中二訳. ファジィ情報学. 日刊工業新聞社, 1993.
- [6] L. A. Zadeh. Similarity relations and fuzzy ordering. Inf. sci, Vol.3, 1971.
- [7] S. Miyamoto. Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer, Dordrecht, the Netherlands, 1990.
- [8] 伊藤哲郎. 情報検索, volume 19 of ソフトウェア講座, 昭晃堂, 1986.
- [9] 浦本直彦. コーパスに基づくシソーラス-統計情報を用いた既存のシソーラスへの未知語の配置. 情報処理学会論文誌, Vol.37 No.12, 1996.
- [10] 益岡隆志 and 田窪行則. 基礎日本語文法. くろしお出版, 改訂版 edition, May 1992.
- [11] 日本電子化辞書研究所. EDR 電子化辞書 1.5 版. 日本電子化辞書研究所, 1995.
- [12] 国立国語研究所. 分類語彙表. 秀英出版, Feb. 1993.
- [13] 中渡瀬 秀一. 統計的手法による分かち書き境界の獲得. 信学技法[言語理解とコミュニケーション] NLC95-77, Vol.96, Mar. 1996.
- [14] 新納浩幸. コーパスを利用した分類語彙表の未登録語義の発見. 情報処理学会論文誌, Vol.38 No.5, 1997.
- [15] 新納浩幸 and 井佐原均. 疑似 n グラムを用いた助詞的定型表現の自動抽出. 情報処理学会論文誌, Vol.36 No.1, Jan. 1995.
- [16] 藤本太郎 and 菅野道夫. 出現情報を用いた新聞記事ジャンルの推定システム. 言語処理学会 第 4 回年次大会 予稿集, Mar. 1998.
- [17] ファジィ学会 編. 講座 ファジィ 9 ファジィデータベースと情報検索. 日刊工業新聞社, Sep. 1993.
- [18] 森本逞 北研二, 小倉健太郎 and 矢野米雄. 仕事量基準を用いたコーパスからの定型表現の自動抽出. 情報処理学会論文誌, Vol.34 No.9, Sep. 1993.

(1998年 7月30日 受付)

(1998年12月10日 再受付)

[問い合わせ先]

〒814-8588

福岡市早良区百道浜2-2-1

㈱富士通研究所 知能システム研究部

藤本 太郎

TEL : 092-852-3261

FAX : 092-852-3272

E-mail : tarof@flab.fujitsu.co.jp

著者紹介

藤本 太郎 (ふじもと たろう)

㈱富士通研究所 知能システム研究部

1993年 広島大学工学部第1類(機械系)卒。1995年 東京工業大学大学院総合理工学研究科システム科学専攻修士課程修了。1999年 同博士課程課程修了。1998年 富士通㈱に入社し㈱富士通研究所に出向、現在に至る。ソフトコンピューティング、自然言語、ナレッジマネジメントに興味を持つ。博士(工学)。

菅野 道夫 (すげの みちお)

東京工業大学大学院総合理工学研究科知能システム科学専攻

1962年 東京大学理学部物理学科卒。同年、三菱原子力工業㈱入社。1965年 東京工業大学工学部制御工学科助手。1977年 同大学大学院総合理工学研究科助教授。1985年 同教授、現在に至る。1991年から1993年の間 ファジィ学会会長。1997年から 国際ファジィシステム学会会長。工学博士。名誉博士。

A Study on the Construction of Verb Thesauri using the Fuzzy Proximity Relation

by

Taro FUJIMOTO and Michio SUGENO

Abstract :

Thesaurus is necessary for natural language processing, however, the construction of the thesaurus requires extreme costs. Hence, several automatic thesaurus construction methods are proposed to reduce these costs. One typical method uses mutual information and this method can construct admissible capability thesaurus, however, it is so complex that it needs large calculation cost. In this paper, we propose low cost and accurate method using fuzzy proximity relation and fuzzy filters.

We use the fuzzy proximity relation to define proximity between verbs, two types of fuzzy filters to reflect appearance information. One fuzzy filter reduces the effect of low frequency verbs and the other reduces the effect of high frequency nouns. We combine two types of filters and fuzzy proximity relations and define four types of fuzzy proximity.

To evaluate these proposal proximity, we use mutual information to define proximity between verbs, and combine two types of fuzzy filters with this proximity. Five types of frequency data characterized by post-position *WO*, *NI*, *HA*, *GA*, *DE* are extracted from EDR co-occurrence dictionary as a training data. First, we re-classify the verbs that are entries of *BUNRUI GOI HYOU* (a Japanese thesaurus) to evaluate the eight proximity. Second, we classify verbs which are not entries of *BUNRUI GOI HYOU* using these eight proximity and evaluate by person.

The result of these two evaluation is that the proximity using fuzzy proximity relation with fuzzy filter which reduces the effect of low frequency verbs using the data including the post-position *DE* gets the highest evaluation. And this fuzzy proximity reduces 7% of calculation time than the proximity using mutual information method.

Contact Address : **Taro FUJIMOTO**

Intelligent Systems Laboratory, Fujitsu Laboratories Ltd.

Fujitsu Kyushu R&D Center, Momochihama, Sawara-ku, Fukuoka 814-8588, Japan

TEL : 092-852-3261

FAX : 092-852-3272

E-mail : tarof@flab.fujitsu.co.jp